Response to the reviewers' comments on "Initial assessment of a multi-model approach to spring flood forecasting in Sweden" by J. Olsson, C. B. Uvo, K. Foster, and W. Yang

First of all we wish to sincerely thank the reviewers for carefully reading the manuscript and for providing many good suggestions for how to improve it. We have implemented the suggested modifications with few exceptions, below follows detailed response to each comment.

General

The manuscript has been turned into a Technical note with the main aim of outlining a possible strategy to improve spring flood forecasts and demonstrate it in a limited case study. The main changes can be summarised as follows:

- Descriptions of individual methods have been moved to a Supplement
- The results for rivers Ångermanälven and Ljusnan have been removed
- Some new text and figures have been included in light of the reviewers comments

Please find detailed response below.

Peter Krahe

R1.1 Please clarify the wording lead times. Starting in the abstract "lead times between 0 and 4 months" are mentioned. E.g. on Page 6082 and later on forecast times for issuing the forecast are mentioned as 1st of January, 1st of March and 1st of May and three monthly volumes of runoff (SFV) are calculated for each of the months.

A figure has been made for clarification, see comment R2.6 below.

R1.2 Only eleven years (2000-2011) are available and used for the comparative analysis of methods, which provides only a small statistical basis.

We fully agree, this is emphasised in the Concluding remarks.

R1.3 By applying the climatological ensemble (CE) I miss the discussion about stationarity in observed climatic and hydrological time series. I expect that also in swede river basins over last 40 to 50 years temperature has been changed, eventually precipitation, too. This can be attributed partly to global climate change. Therefore, is it still allowed to use e.g. data of 1961 to 1990 for actual seasonal forecasting? This issue has to be discussed a bit. This is also related to new variants of the analogue ensemble method (AE).

This issue is of course very relevant and should indeed have been included in the original manuscript, that we assume climatic stationarity in the AE approach. Any clear trends are, however, not obvious in the historical observations and it has previously been shown that using a more recent period in the CE method does not improve performance as compared with using the full historical period. We have included a discussion in the Supplement (page 2).

R1.3 In Table 2 where the results of CE are depicted I miss the MAE's for the simulated values. Only one value for SIM is listed of which it is not clear which period is captured. I expect that MAE will strongly depend from the analysed periods. Therefore, MAESim 1/1, . . .SIM1/3 and . . .Sim1/5 are of interest.

These values were unfortunately not properly explained in the original manuscript. In the table (Table 1 in the revised manuscript) SIM refers to SFV simulations with observed P and T ("perfect prognosis") and F refers to CE forecasts. This has been clarified.

R1.4 Within this short evaluation period inhomogeneities in the seasonal forecasts are present, e.g. the change from ERA40 to ERA 15 (2003), Meteorological ensemble consisting of ECMWF IFS-Hope with 11 members up to 2006 and 41 members later on, the role and the range of ARPEGE-ORCA is unclear (this model is only used in SD and only 1 member is available ?).

These inhomogeneities are certainly a limitation, we fully agree. We have assumed that their influence is small, and this has been added in the revised manuscript (Concluding remarks). About ARPEGE-ORCA model, this is used only in the SD method, yes, and with an ensemble similar to IFS, this has been added.

R1.5 Application and discussing the Analogue Ensemble (AE), is based on period 1961-1999 and station data while gridded data are used for hydrological modelling (calibration and validation ?, cp. 3.2 versus cp.2.1)

The gridded data are optimized for hydrological modelling, by undercatch correction, optimal interpolation, etc. For the CP classification, however, station data are usually preferred and used .

R1.6 The daily precipitation and temperature data of the ECMWF-IFS model is used, but no bias correction is applied even the necessity to do this is known by the authors (see cp.1, p. 6079 and 6080), at least in the context of climate models. Furthermore, the mapping process to downscale the raw output to the grid and/or station location required by the hydrological models is not well described.

We chose not to include any bias correction in this initial test. Actually ongoing work indicates that the positive impact of bias correction on seasonal forecasts in Vindelälven is rather small. And although bias correction generally improves overall performance it is not free of problems, e.g. commonly used methods may modify extremes in unwanted ways and it needs updating when model versions change, as discussed in the Concluding remarks.

Concerning the mapping process, this has been further explained in the Supplement.

R1.7 Concerning the dynamical meteorological models (DM and SD) it has to be noted that the analysed system (at least ECMWF system 3) is replaced by system 4 since 2011. Furthermore, hindcasts exist for a 20 year period for this system. Therefore, the findings based on system 3 are of limited value. However quotations should be added concerning system 3 on p.6083 and on p. 6097, assuming system 4 of ECMWF is the new system which becoming available for SMHI?

Yes, this is clearly a limitation. The work presented in the paper was carried out over an extended period of time and we have addmittedly been a bit slow with moving from System 3 to 4. The system is currently been updated and re-evaluated using System 4 data, snow input, more catchments, longer series, etc., but this will take a while. We have done some limited testing of the impact of system version on the DM results and it actually appears rather limited, not any drastic overall improvement. This is included in the revised manuscript (Concluding remarks).

R1.8 Due to the mentioned restrictions concerning the application and interpretation of the dynamical models, both the DM and SD, the results should be interpreted carefully. Please take more care for this in the discussion.

We have further emphasised the limitations in the Concluding remarks.

R1.9 I see a limitation by applying the SD-Method without considering the previous period, e.g. the snow accumulation in the wintertime is not taken into account. This is although reflected in the results of Table 3 for forecast times 1/3 and 1/5. On the other side you have the best result of all methods for SD and forecast time 1/1. This issue should be discussed a bit more.

Snow accumulation is being implemented in an updated version of the system, and this discussion is included in Concluding remarks.

R1.10 For better readability in cp.3 p.6098 the first paragraph concerning the predictors should be depicted in a table.

Done, Table S2 in the Supplement.

R1.11 By applying the CP approach the criteria for selection are the two most frequently occurring CPs in the previous months 1 to 6. What is justification for selecting 2 and e.g. not 3 or 4 CP's? Please, explain a bit more.

We tried different combination and found that more CPs than two did not add any skill in the SFV forecasts, this has been included in the revised manuscript.

R1.12 It may be an added value as well as it will support the interpretation of the results if you give some more information of occurred CPs in cp.5.1.2 in form of a table or a histogram, e.g. for CP3. It may be interesting to see e.g. for a certain forecast time that always the same CP's dominate the 2 leading CP's or not.

We agree this is potentially interesting but as the manuscript has been turned into a Technical note we generally prefer not to go into any more detail on the models than in the original manuscript.

R1.13 Concerning hydrological modelling/processes the following remarks has to be noted:

R1.13a What are the catchment sizes of the sub-basins of the HBV Model, e.g. add in Table 1 number of sub-basins.

To save space we have merged Tables 1 and 2 in the original manuscript into Table 1 in the revised manuscript, and this table is essentially "full". The additional information requested here (and in comments R1.13c and R1.13d below) have been added in text.

R1.13b Some more words to calibration and validation strategy are needed especially taking into account the water resource management systems in at least two of the rivers. Some remarks about the temporal stationarity of the hydrological modelling may be helpfully.

The operational model used is not temporally stationary in the sense that it is re-calibrated on a semi-regular basis, following evaluations of each year's performance and decided in connection with end-users. It is difficult (if even possible) to give any meaningful numbers on calibration/validation performance, but we can only say that it is supposed to be the best available model and give the overall perfomance measures in Table 1. We have clarified this.

R1.13c What is the role of natural lakes in the system. E.g. characterize the lakes by their number, total area or volume in Table 1.

See comment R1.13a.

R1.13d Furthermore, the mean catchment elevation as well as some basic climatic data characterizing the catchments (e.g. mean annual or seasonal precipitation and air temperature) will be helpful.

See comment R1.13a.

R1.13e Hydrographs characterizing the streamflow regime in the considered seasons will be helpful (e.g. 31-day moving average of daily runoff of the three basins). Although it may be interesting if there are changes in the precipitation –runoff process are obvious in the observed data.

Figure 2a with mean hydrograph and associated text has been added.

R1.13f The results depicted in Table 3 show a large variety between the river basins. Therefore, it may be interesting to calculate and to show the correlation matrix between the river basins for the various forecast periods.

Only the results for Vindelälven are kept in revised manuscript.

Some technical/linguistic remarks:

R1.14 A table with an overview of the applied methods and data including the required temporal resolution will be helpful.

To save space we have chosen not to include any table, but we hope that the additional information provided in response to other comments will suffice

R1.15 Figure 1is not very illustrative at time. Due to the fact that the seasonal forecasting methods, especially CP and SD require information from a larger area, these areas should be depicted in the map.

The CP and SD domains have been included in Figure 1.

R1.16 p. 6080, line17: skip "in" for impacts"

Skipped.

R1.17 p. 6082, line 14: change to 900 x 106

Changed.

Massimiliano Zappa

R2.1 I have a general concern with respect to the title of the manuscript. Why multi-model? HBV is the only model, you have only multiple approaches for selecting and/or generating your forcing data. Why flood? You are just looking at the cumulated discharge volume during the snowmelt-season. I guess I would have called the study: “Initial assessment of spring discharge volume forecasting in Sweden: evaluation of multiple strategies to select atmospheric forcing” I admit, that this is less fashionable than your original title, but I am convinced, that this is a better choice to declare the work you present.

Multi-model: Well, we use only one *hydrological* model but we also use atmospheric models and statistical models. But OK, it may be misleading and we have changed to the more general “multi-method”.

Flood: To us “spring flood” is an established term for the spring snow-melt runoff peak (even if not causing flooding). It used by e.g. NOAA/NWS (http://www.weather.gov/nerfc/springfloodpotential) and we would prefer to keep it.

Issues to be addressed (Page(s) – Line(s)):

General comments:

R2.2 You present here several ways to upgrade your seasonal predictions, all of them focussed on the meteorological input. Are there any reason for ignoring for instance the initial conditions of the model (e.g. selecting years with similar snow water equivalent at the beginning of the forecast period)?

The initial conditions is certainly important, and we work a lot on that too, but in this study we limit ourselves to investigating the impact of the meteorological input, assuming that the simulated initial state is representing the best we can attain.

R2.3 As well pointed out by Peter Krahe in his review, many of the data you are using are affected by instationarity, lack of homogeneity and biases. For instance the use of raw output of IFS (and of any numerical model) is something you really need to focus on. We recently presented two studies demonstrating the value of calibrated meterological data (Fundel and Zappa , 2011; Jörg-Hess et al., 2015).

We fully agree, see response to comments R1.4 and R1.6 above.

R2.4 6082 – 10-12: You declare here “In this study we focus on forecasts of the accumulated discharge in the spring flood period (May–July)”. In my opinion the accumulated discharge during the flood period cannot be declared as “spring flood forecast”. I think it would be more adequate to declare your product as “spring discharge volume”.

We disagree, see response to comment R2.1 above.

R2.5 6083 – 13-17: As I understand here you declare the outcomes of model calibration and also present them in Table 1. I would expect here also some words and numbers on model verification.

See response to comment R1.13b above.

R2.6 6092 – 2 – 9: I come here a little bit in trouble with respect to your definition of lead time. I think you should introduce a second figure presenting the setup concerning lead time and evaluation period. E.g.:

Yellow: Model running

Blue: Model running and evaluation of discharge volume.

Good idea, a figure along these lines has been included (Figure 3).

R2.7 6100: The weighted multi model is of course an interesting section of this paper. I really like the simple way you present here to assign the weights (based on ranking). Is there a publication you can cite that presents an assessment of this for me very elegant approach? Do you plan to evaluate other methods to assign the weights (e.g. Bayesian averaging)?

Some reference conceivably exists but we do not know of any. We are indeed studying also other more sophisticated weighting approaches but with only the limited period used in this study, nothing more advanced than this simple method appeared meningful.

Minor comments:

R2.8 General issue: I'd like to learn more about the variable you are forecasting. How large is the variability of these spring discharge volumes? Could you add the minimum and maximum discharge volume (May-July) within all years of your experiments in Table 1?

Table 1 has been updated with min/max and the new Figure 2b shows the climatological SFV distribution.

R2.9 6080 – 22 to 6081 - 9: This paragraph sounds more like a methodological section. Is there any way to formulate this more generally to introduce the goals of your study and move the present formulation in the methods section (e.g. at the beginning of page 6085)?

This paragraph was admittedly misplaced. It has been deleted and parts have been removed to more suitable positions in the revised manuscript.

Final considerations:

R2.10 I am struggling with the recommendation for this manuscript. I acknowledge the value of these preliminary results and I can also support the fact that you present the data as they are without any post-processing, because this is the starting point and one should be aware of this. I think this manuscript would profit to a “demotion” to “technical-note” with a title sounding like “Technical-note: Spring discharge volume forecasting in Sweden - Evaluation of multiple strategies to select atmospheric forcing”. I recommend therefore major revision to re-shape the manuscript to be published as technical note.

It is a reasonable suggestion that we have followed. By mainly (1) moving the methodological descriptions to a Supplement and (2) omitting the results for rivers Ljusnan and Ångermanälven, we believe the format is suitable for a Technical note (even with the additions required in light of the comments).

Renaud Marty

R3.1 A number of assumption and details about dynamical and statistical approaches need to be justified and clarified. Some of them came from literature. Thus it is quite difficult to judge the validity and the contribution of the authors.

We have tried to make our own contributions as well as their validity more clear, see the response to several other comments above and below.

R3.2 To my opinion, the verification part is subject to sample effect and the manuscript is confused on this topic. Especially it is not clear what are the ensemble sizes provided by the different procedures.

Ensemble sizes were admittedly not properly reported in the original manuscript. In the revised manuscript they are included in the Supplement (Table S1).

R3.3 The authors should explain why a spring flood forecast is reduced to a deterministic forecast from an ensemble of possibilities.

We fully agree that there is an added value in ensembles and in the multi-model (now: multi-method) we do compose an ensemble represented by one member from each method (i.e. maximum five). However, we do not find probabilistic evaluation meningful with this small sample (11 years) but only of the median and a weighted mean.

But it is true that each method is reduced from an ensemble to one deterministic forecast. As written above, each method had its own ensemble size and pooling ensembles of different size is not a trivial matter, especially as some methods produced a different size from year to year. This is an issue we are currently struggling with and hopefully have a solution for soon. We have added a discussion (page 8) together with giving the ensemble sizes (see reply to R3.2).

R3.4 Why the authors kept only 3 forecast issue dates? The results will be more conclusive with a greater sample.

We fully agree. We restricted ourselves to three dates (early/middle/late) for practical reasons and in retrospect we should definitely have used more.

Furthermore they are specific comments that the authors should consider to improve their manuscript. Note that I don't have a wide access to the majority of papers cited in this article. This review does not concern the references.

Specific comments

R3.5 In page 6083, line 6: “Applying the model calibration of a number of free parameters, generally aout 10.” This part requires more details, including the explanation of having several parametrization of the same model. Do the 3 basins have the same parametrization, i.e. the same number of free parameters?

The parameterisation (calibration) was different for each basin, but in the revised manuscript only Vindelälven is kept.

R3.6 In page 6084: why do you describe the ECMWF's model in details (“Cy31r1”) and not the Arpege one? This level of details is unnecessary in your manuscript

It has been removed.

R3.7 In page 6084: Are ECMWF and ARPEGE models used in the same time or as two configurations of SD approach?

Each model produces one forecast ensemble, which are pooled into one “SD ensemble”. This description has been included in the Supplement.

R3.8 In page 6085: The description of the CE procedure is clear. Nevertheless, could you give the CE ensemble size? Please explain why the simulations are not made from the start of the last hydrogical year?

The CE is made up of all historical years from 1961-‘present’, this means that the CE has 40 members in 2000 and increases in size by one member for each year thereafter. The spin up period is from 01-01-1961 to ‘present’. As each new forecast is made the initial conditions, or model state, are saved and these are used as the initial conditions when spin-up for the next forecast date are performed. This has been added in the Supplement.

R3.9 In page 6086: The relation between indices and hydro-meteorological trends is clearly described but deserves some figures to facilitate the analysis

We are uncertain what kind of figure the reviewer has in mind here. We have expanded the description somewhat for further clarification.

R3.10 In page 6087: Some information seems to be implicit in TCI procedure. Please clarify that the period length (1 to 6 months) is a parameter. The configuration TC1 determines the analogue year with the comparison of 3 value per year, while TC6 needs the concordance over 18 values per year. The interpretation of the ongoing results may be submitted to sample effect. Especially when you say "If not analogue years can be identified among the historical ones by comparison of the state of the three climate indices, analogue years are sought using an agreement with two of them.". Is there a indice that is always remove from the comparison sample? How many TCI members make your sample? Or the analogue year is the input to provide your median SFV? Please clarify.

We have expanded this paragraph to clarify the TCI procedure, see the Supplement (section S2.1). See also reply to comment R3.2 above.

R3.11 In page 6088: Why the CP procedure is calibrated only for the Vindelälven basin? Why CP is not set for the other ones?

It was mainly due to limited resources (and that the same classification worked well in both of the northern rivers), but in the revised manuscript we only use Vindelälven

R3.12 In page 6089: In you Eqn (3), d seems to be a parameter that can be optimized in CP procedure. Do you use 0.1mm as mentionned in your manuscript to be the general case or did you use another threshold?

We used 0.1 mm as mentioned in the manuscript, this has been clarified.

R3.13 In page 6089, line 12: "... are determined subjectively to adjust for differences in magnitude as well as importance" sound not clear. What do you mean by magnitude? Spread? Please clarify.

"Magnitude" referred to the relative magnitudes of O1 and O2 (which may be compensated by the weights), this has been added.

R3.14 In page 6089, lines 16-17: "CP catalogue" is not defined. Is there a set of weather types?

The sentence referred only to the circulation patterns themselves; the word "catalogue" was confusing and has been removed.

R3.15 In page 6089, lines 20-22: "The two mos frequently occuring Cps within a period of 1 up to 6 months prior to the forecast issue date are used as a criterion to select the analogue historical years" Do you mean that historical CP series are compared to these two CPs? Finally, how many analogue years constitute your sample?

The occurrence frequency of each CP catalogue is calculated from historical and forecasting CP time series. The two most frequently occurring CPs are compared. Thus, selected analogue years are different from year to year. See also reply to comment R3.2 above.

R3.16 In page 6090, lines 10-11: Your assumption is strong (see e.g. Buizza and Palmer, 2008, Monthly Weather Review, 126), especially for precipitation ensemble forecast. Please justify that the ensemble median is not influenced by the ensemble size, increasing from 11 up to 41 in your dataset.

As written above (R3.2-3.3) we are dealing with many different ensemble sizes and we have used the pragmatic and consistent approach of using the median in our forecasts. We have clarified this and included a reference to the paper by Buizza and Palmer.

R3.17 In page 6091: As the SD method is stochastic, how is defined the ensemble median?

The three best predictors from the atmospheric models are used to form an ensemble, from which the median was calculated. This was not clear in the original manuscript but has now been added in the Supplement.

R3.18 In page 6092: The evaluation is only based on 3 forecast issue dates. The paper should present more forecasts. e.g. one per decade, to outline the influence of the lead time.

See response to R3.4 above.

R3.19 In page 6092, lines 8-9: "They were selected by an initial screening period based on previous literature..." You must give the reference of this previous literature.

References have been included.

R3.20 In page 6092, lines 20-21: You could indicate that an approach is perfect if RI is 100%.

Done.

R3.21 In page 6094, line 10: Which sample do your refer by "limited sample"?

We meant only 10 years of forecasts. But this text concerned river Ljusnan and has been removed in the revised manuscript.

R3.22 In page 6095, line 4: According to the results shown in table 2, the first statement is wrong. Please clarify.

We meant that the forecast results (in terms of RI and FY) were overall similar, not necessarily the historical simulations, this was not admittedly unclear. In the revised manuscript we show results for both stations in Vindelälven

R3.23 In pages 6096-6097: To be more robust your analysis deserves to include CP optimized basin by basin, and globally. Or please justify your decision to optimize CP only on Vindelälveln basin.

See response to R3.11 above.

R3.24 In page 6097, lines 23-27: It would be interesting to include the new version of ECMWF seasonal forecasting in your results

See response to R1.7 above.

R3.25 In page 6098: Is there a reason to justify a SD configuration per forecast issue date? Do a global configuration decrease the performance of SD approach? In Section 2 the meteorological dataset is made of two atmospheric and oceanic combination. Is SD procedure applied on both combinations or only on the one including the ECMWF model? Is SD equally sensitive to the ECMWF or ARPEGE ?

Yes, a global configuration of the SD approach negatively affects the performance. The reason for this is that the relationship between the predictors and the evolution of the snowpack differs over the season. The SD approach is applied to both combinations. We did not perform any analysis of how sensitive the SD approach was to the different data streams, however it was noted that variables from the ARPEGE data stream was represented more often in the SD multi-model than those from the ECMWF data stream. See also response to R3.17 above.

R3.26 In page 6099, lines 9-12: "If any of the new methods could not generate any forecast, it was replaced by CE [...]" Does it mean that previous results shown is Section 5 are based on heterogeneous sample of deterministic forecasts? If that is the case, it isn't mentioned anywhere and limit the scope of the previous conclusions. Please clarify and please give the importance of such replacements.

In a very few cases the CP method, as implemented here, was not able to identify any analogue year and then it was replaced by the CE forecast to have a complete time series of forecasts. It has a negligible impact on the results. We have moved this text to the description of the CP method in the Supplement.

R3.27 In page 6099, lines 14-17: The statement sounds explain that the median is less sensitive to sample issue. This aspect is too implicit in the manuscript and need to be addressed more clearly in Sections 3 and 5.

We have changed the formulation. See also response to R3.16 above.

Technical corrections

R3.28 In page 6078, line 2: what means "useful" in your context?

Changed to "accurate".

R3.29 In page 6080, lines 11 and 15: "teleconnections" is not widely used in statistics. Could you give a technical definition of this term?

We have expanded the explanation.

R3.30 In page 6081, line 19: "development" could have a double meaning here. Do you refer to meteorological situations?

Yes, but changed to "weather conditions".

R3.31 In page 6082, Eqn. 1: what don't you use a symbol to represent "lakes"?

For some reason this is an established way of writing this equation. But OK, "lakes" has been changed to "VL" in the revised manuscript.

R3.32 In page 6083, line 14 and hereafter: you should use NSE instead of R2 to represent the NashSutcliffe efficiency. R2 is more commonly used as the determination criteria in statistics.

Changed.

R3.33 In page 6083, line 24: you should put the CPC's domain name in your references and, here, give only the reference (e.g. CPC)

Changed.

R3.34 In page 6085, lines 24-25: typo “A collection of … constitutes the historical data”

Changed.

R3.35 In page 6086, line 1: “better” refers to a comparison. Here, the term is incorrect. “accurately” or “correctly” are more appropriate.

Changed.

R3.36 In page 6086, lines 16-18: Please, move “, which is part of … (NOAA)” in page 6083. The internet link is already given in Page 6083.

Changed.

R3.37 In page 6087, line 22: you should replace “normally” by “generally”

Changed.

R3.38 In page 6091, lines 24-25: In hydrological forecasting, lead time is sometimes defined as the period between the forecast issue date and the validation date of the current forecast.

We have illustrated our definition of lead time in the new Figure 3.

R3.39 In page 6092, line 17: type “MAE_{CE}”

Changed.

R3.40 In page 6094, lines15-19: This part has to be moved in conclusion of Section 3

This text has been moved to Section 3.

R3.41 In pages 6094-6095, lines 27-28 and 1-2: theses results need to be added to a Table

This part has been removed.

R3.42 In page 6098, last paragraph: The statement would be more comprehensible with an illustration

We have difficulties interpreting what type of illustration the reviewer has in mind, it is a rather general statement.

R3.43 In page 6102, line 5: “and will be reported elsewhere” is unnecessary

Removed.

R3.44 In page 6113, Fig.1 : Is Vindeln the same basin known in the manuscript as Vindelälven ? Please keep one denomination.

Changed.

Philippe Crochet

Major comments:

R4.1 A few additional figures would be welcome to help analysing the results. Showing the distribution of SFV and some statistics (min/median/max) would help understanding how large is the inter-annual variability on the different basins. Scatter plots of forecast vs. observed SFV, or time series plots would help the reader to better understand and compare the quality of the different SFV forecasting methods. This should also help analysing some discrepancies in the results, when e.g. FY>=50% while RI is negative (See Table 3). What makes RI negative in these cases ? a few outlier years ? Can we conclude that the new methods in question are worse than the reference one in these cases ? Which criteria is the most important in the present context: RI or FY ?

We have added SFV distribution in Figure 2 and min/max-values in Table 1. Scatter plots are indeed informative but 1/ they are very different from case to case (method, forecast date, station…) and it hard to say what is most representative and 2/ we have limited space in a Technical note. So we have not added any scatter plot. The issue concerning RI and FY, their sensitivities and their importance, is indeed interesting and relevant and we have added a discussion in the revised manuscript.

R4.2 It appears that HBV has been calibrated on the period used to evaluate the forecasts rather than on an independent period (Table 1). The skill of the hydrological model is most likely overestimated, compared to what it would be in an operational environment. This approach could be acceptable if HBV-based methods only were compared in the study. However, in the present case, I would think that the SD method is disadvantaged. This should be discussed at least, if not reconsidered.

Table 1 shows only the performance in the evaluation period, it does not imply anything about the calibration. The HBV model is the operational version, and this has been added (see also response to R1.13b above). The calibration is mainly based on the historical period prior to the evaluation period (1961-1999), but some re-calibration has been done also later. The SD method was calibrated for the period 1982-1999. Thus the SD method may indeed be slightly disadvantaged but we believe the impact is small. We have clarified the calibration strategies in the revised manuscript.

R4.3 The use of ERA40 and ERA-Interim to run the CP-based analogue method in forecasting mode is questionable. The present application of the method does not reflect the expected skills in operational conditions and the improvements relative to the baseline method are possibly overestimated. In an operational environment, operational NWP analyses rather than ERA-40 or ERA-Interim will be used to define the CPs to be compared to historical CPs derived from ERA-40 and ERA-Interim. Therefore, I would expect that discrepancies or inconsistencies between ERA-40 and operational NWP analysis will impact the method's skills. In order to better reflect expected shortcomings in operational conditions, I would recommend to either re-calculate the anomaly g(i,t) and corresponding CPs for the years to be forecasted, using available operational NWP data, or change the validation period so that ERA-Interim only is used in forecasting mode and compared to ERA-40 historical CPs.

Indeed, the CP method as implemented here is not directly applicable in an operational setting, but the results should be viewed as only theoretically attainable. Unfortunately we are not able to re-calculate as suggested during this revision, but instead we have 1/ clarified this limitation of the CP approach and 2/ included two ensembles in the final multi-method, one taking all new approaches into account and one taking all operationally available ones into account.

R4.4 Since three of the new methods (two analogue methods and ECMWF seasonal forecast) aim at improving the meteorological forecasts, they should have been evaluated with that respect before being evaluated with respect to SFV.

Such evaluation has been performed also for the analogue methods, but was not included as the SFV results overall well reflect how accurately the analogue years are able to represent the actual years in terms of meteorological conditions. In the revised manuscript we have omitted also the evaluation of the daily ECMWF forecasts 1/ to be consistent with the analogue approach, 2/ because System 3 is outdated and 3/ as a part of the compaction of the manuscript into a Technical note.

R4.5 In particular, considering the coarse horizontal resolution of ECMWF seasonal forecasts, biases can be expected to affect precipitation and temperature forecasts at the basin scale, especially in complex terrain. This should have been investigated before using them into HBV rather than after. A bias correction or an adaptation strategy is probably necessary. On the other hand, the frequent model updates can be an obstacle to the development of robust correction strategies. This should also be discussed.

See reply to comment R4.4 above. We have expanded the discussion in Concluding remarks about pros and cons of bias correction.

R4.6 The strength and limitations of analogue methods in general and the proposed ones in particular need to be discussed. What about: i) the impact of the archive length on the correct identification of analogue years, ii) the impact of the number of selected analogue years on the quality of the reduced ensemble and its median value, iii) the validity of the assumptions behind the proposed analogue methods, i.e. the degree to which antecedent meteorological conditions, 1 to 6 months prior to the forecast issuing date, are relevant to the prediction of future meteorological conditions up to springtime, iv) the validity of the optimisation method for CPs, which is based on precipitation only and not temperature, yet a key variable in the formation and melting of snowpack. Also, why is the optimisation conducted on one catchment only ?

i) The archive length is indeed a limitation; with more years available the possibility of good analogues being present increases. We have designed the two analogue approaches somewhat differently, with TCI always finding an analogue but not CP (using CE as fallback). We have expanded the introduction of analogue approaches in the Supplement (section S2).
ii) See response to comment R3.16 above (as well as other comments by Reviewer 3).
iii) See response to comment R3.10 above (as well as other comments by Reviewer 3).
iv) Temperature is indeed important, but we considered precipitation to be the most important control on SFV. Including also temperature would reduce the probability of finding distinct analogues (see also i) above). Only the catchment on which optimisation was conducted is included in the revised manuscript.

R4.7 Concerning the SD method, it seems that forecasts (rather than an analysis) are used to define the predictors for the calibration of the predictors-predictant relationships. The forecasting errors may deteriorate the nature of the underlying relationships. What is the rationale behind that ?

SD simply establishes a statistical relationship among the predictors and predictand. The physical underlying relation is not taken into consideration. This is actually the strength of the SD, it "corrects" the weaknesses of the models.

R4.8 I was wondering whether the skill of the multi-model approach could be further improved after eliminating the worst method(s) and/or by defining a weight proportional to the skill of the method rather than just the rank. I am also a bit surprised to see that the multi-model approach can perform better than the baseline method, when all new methods seem to perform worst.

We have changed the multi-method approach and excluded the worst performing analogue method. With a long evaluation period available skill-based weighting may be better but with only 11 years available it results in overfitting. In some cases, the multi-method indeed performs slightly better than each of the single methods included. Generally, this is because very inaccurate single forecasts become eliminated, and this has been included in the revised manuscript.

R4.9 Ensemble forecasts are developed but only the ensemble median is evaluated. The advantage of making use of ensembles is not discussed. In particular, ensemble forecasts should be reliable.

See response to R3.3 above.

Specific comments

R4.10 p. 6082: lines 7-8: is the regulation of Angermanalven and Ljusnan rivers taken into consideration by HBV ? Does this have an influence on the skills of the different methods, including those not using HBV ? This should be kept in mind when analysing the different results.

The impact of regulation is a complicated issue that we are currently looking further into. The two regulated rivers have been excluded in the revised manuscript.

R4.11 p. 6082 line 14: do you mean 900 x 10ˆ6 and 8000 x 10ˆ6 ?

Yes. We have changed this paragraph.

R4.12 p. 6083 line 1: I would suggest to write somewhere that the output of HBV is daily Q, although this might be obvious to those familiar with HBV.

Added.

R4.13 p. 6085, lines 14-15: Are you using all historical time series to run the CE baseline method, including the current year under investigation, or only those prior to current hydrological year ?

Prior, this has been added.

R4.14 p.6087 line 6: can you explain how is persistence defined and used ? Are the TCI calculated for individual months or for the entire period (1 to 6 months) ?

See response to comment R3.10 above.

R4.15 p. 6087 lines 11-15: have you tried to use one TCI only rather than a combination of 2 or 3 TCI's ? what is the rationale behind this choice ? is it the result of an optimisation ?

See response to comment R3.10 above.

R4.16 p. 6087 (TCI) and p. 6089 (CP?s): the number of selected analogue years differs from year to year for each method and lead times. How is this impacting the quality of the ensemble and resulting median ? Please give the number of selected analogue years for each forecast year in a table or at least some statistics (min, max, median). If too many years are selected, compared to the total number of years, then the method will converge toward the climate ensemble baseline method. If too few years are selected, the uncertainty of the reduced ensemble will be very large.

We have included statistics of the number of identified analogue years (Table S1). Indeed, the number of years used has impacts as suggested. We have not been able to make any more in-depth analysis during the revision but we have expanded the text in the revised manuscript.

R4.17 p.6088: how do you go from Eq. 2 to CPs? Do you define several g(i,t) classes ? Please explain better.

The Eq. 2 is used to calculate the predictor, g(i,t) , for the CP classification. g(i,t) indicates the deviation of daily MSLP from the long-term climatology. Every g(I,t) is categorized to one of 5 groups using fuzzy logic: large positive deviation, relatively large positive deviation, no obvious deviation and relatively large negative deviation and large negative deviation. To determine the best possible rule sets, fuzzy rules have to be optimized with a local variable using an objective function that explains variability in frequency and amount of precipitation in study area. Thus, each CP can be finally described with a fuzzy rule k, represented by a vector V(k) = (v(1)k, v(2)k, . . . , v(i)k, i=1, n). Here, n is the number of locations, g(I,t), and k stands for the CP. This description has been included in the Supplement.

R4.18 p. 6089 line 18: how many CPs have you defined ? Can you give statistics on the frequency of occurrence of each CP over all years ?

See response to R1.12 above.

R4.19 p. 6089 line 19: how do you define persistence and how is it used ?

In practice persistence was never used in this implementation of the approach and the word has been removed.

R4.20 p. 6089 line 20: did you arbitrarily define the rule of using the 2 most frequently occurring CPs or did you try other possibilities ? did you validate this choice against forecasted P and T ? Please explain better.

See response to R1.11 above.

R4.21 p.6092 lines 4-9: I would suggest to move this paragraph to Section 3.4. Also, a figure illustrating the practical application of the statistical downscaling method would be appreciated, such as a scatter plot of the relationship(s) in calibration mode (median prediction against observed SFV).

The paragraph has been moved. Concerning a scatter plot, see response to R4.1 above.

R4.22 p.6092: lines 4-5: If I understand correctly, the seasonal predictors are calculated for different periods. For a forecast issued on 1 January, the predictors are defined for the period Jan-Feb-Mar, for a forecast issued on 1 March, the predictors are defined for the period Mar-Apr-May and for a forecast issued on 1 May, the predictors are defined for the period May-Jun-July ?

Correct, this has been illustrated in the new Figure 3.

R4.23 p. 6093 line 10: According to Eq. 6, MAE is defined in m3 and not in %. I would suggest to either redefine MAE in Eq. 6 to make it in %, or to give results in m3

MAE (now MARE) has been redefined.

R4.24 p. 6094: line 23-26: A direct assessment of weather forecast errors will help to clarify the situation.

To some degree, yes.

R4.25 p.6096 lines 1-4: What do you mean by climate phenomena ? can you explain better ?

This text has been removed.

R4.26 p. 6097, line 1-2: This demonstrates the importance of calibrating the method for each catchment separately. Why not doing so ?

It was due to resource constraints. Ljusnan has been removed in revised manuscript.

R4.27 p. 6097, line 3: was it in Section 3.1.2 or 3.2.2 ?

The latter, but this text has now been moved.

R4.28 p.6097: lines 5-10: You claim that a higher uncertainty is observed because discrepancies between ERA-40 and ERAInterim datasets are leading to inconsistencies in the CP classification. Therefore, if the method was applied operationally, operational ECMWF analysis would have to be used to define the CPs for the past 1 to 6 months, and there would be discrepancies too. So in practise, I am not sure that you will be able to obtain this 10-20% improvement.

See response to R4.3 above.

R4.29 p.6097, lines 16-17: this should have been done first.

See response to R4.5 above.

R4.30 p.6098, lines 1-7: I don't quite see the rationale behind the use of each of all these parameters and some explanations would be welcome. Also, I would prefer to see the list of predictors for each forecast issue date.

The rationale is that this are the parameters identified by correlating them to the discharge and forecast issue dates were given. We have updated them to represent only Vindelälven (Table S2).

R4.31 p. 6098, line 16-17: same with analogue methods.

True, it has been added also in their description.

R4.32 p. 6099, line 10-11: This should be mentioned in the presentation of the CP approach.

It has been moved.

R4.33 p. 6099, lines 18-21: why ? could this be related to a few poor forecasts only ? Scatter plots may help understanding this better. Also, add information about FY in Table 6.

See response to R4.1 above. FY has been added in Table 6 (now 3).

R4.34 p. 6109, Table 3: Please give a global average for RI and FY for each method, over all rivers and lead times, as in Tables 4 and 5.

Included.

R4.35 p. 6110: Table 4, TCI6 1/1: results are slightly different than in Table 3. Please check values.

These values are removed in revised manuscript.

R4.36 Tables 4 and 5 could be skipped.

Skipped.

R4.37 p. 6113, Fig. 1: Please add information about the spatial domains used to define CP and SD.

Added.

Technical Note: Initial assessment of a multi-method~~model~~ approach to spring flood forecasting in Sweden

J. Olsson[1], C. B. Uvo[2], K. Foster[1,2] and W. Yang[1]

[1] Research & Development (hydrology), Swedish Meteorological and Hydrological Institute, 601 76 Norrköping, SWEDEN

[2] Department of Water Resources Engineering, Lund University, Box 118, 221 00 Lund, SWEDEN

Correspondence to: J. Olsson (jonas.olsson@smhi.se)

Abstract

Hydropower is a major energy source in Sweden and proper reservoir management prior to the spring flood onset is crucial for optimal production. This requires ~~useful~~ accurate forecasts of the accumulated discharge in the spring flood period (i.e. the spring-flood volume, SFV). Today's SFV forecasts are generated using a model-based climatological ensemble approach, where time series of precipitation and temperature from historical years are used to force a calibrated and initialised set-up of the HBV model. In this study, a number of new approaches to spring flood forecasting, that reflect the latest developments with respect to analysis and modelling on seasonal time scales, are presented and evaluated. Three main approaches, represented by specific methods, are evaluated in SFV hindcasts for ~~three main~~the Swedish river~~s~~ Vindelälven over a 10-year period with lead times between 0 and 4 months. In the first approach, historically analogue years with respect to the climate in the period preceding the spring flood are identified and used to compose a reduced ensemble. In the second, seasonal meteorological ensemble forecasts are used to drive the HBV model over the spring flood period. In the third approach, statistical relationships between SFV and the large-sale atmospheric circulation are used to build forecast models. None of the new approaches consistently outperform the climatological ensemble approach, but for ~~specific locations and lead times~~early forecasts improvements of up to 25~~20-30~~% are found. This potential is reasonably well realised in a multi-method system, which over all forecast dates reduced the error in SFV by ~4%. This improvement is limited but potentially significant for e.g. energy trading. ~~When combining all forecasts in a weighted multi-model approach, a mean improvement over all locations and lead times of nearly 10% was indicated. This demonstrates the potential of the approach and further development and optimisation into an operational system is ongoing.~~

1 Introduction

In Sweden, seasonal (or long-term) hydrological forecasts are used primarily by the hydropower industry for dam regulation and production planning (e.g. Arheimer et al., 2011). The forecasts may be used to optimise the balance between a sufficiently large water volume for optimal power production and a sufficient remaining capacity to safely handle sudden inflows. In northern Sweden, the spring flood forecast is the most important seasonal hydrological forecast and it generally covers the main snowmelt period in May, June and July.

Traditionally, discharge and spring flood forecasting at seasonal time scales have been based on two approaches. The first utilises statistical relationships between accumulated discharge during the forecasting period and predictors such as snow water equivalent and accumulated precipitation that represent the hydrological state at the forecast date (e.g. Garen, 1992; Pagano et al., 2009). The other approach is based on a hydrological model, which is initialised with observed data up to the forecast issue date and then forced with historical meteorological inputs over the forecasting period (e.g. Day, 1985; Franz et al., 2003). In addition, hybrid approaches, applying model-derived information in the statistical regression, have been proposed (e.g. Nilsson et al., 2006; Rosenberg et al., 2011).

Recently, substantial progress has been made in the field of seasonal climate forecasting. It may be distinguished between dynamical and statistical approaches. In the dynamical approach, numerical atmospheric models (global circulation models - GCMs) have been developed to predict seasonal climate, i.e. the average climate for three consecutive months, several months ahead (Goddard et al., 2001). The scientific basis of such predictions is that the sea surface temperature (SST), that characteristically evolves slowly, drives the predictable part of the climate. Consequently, providing to a GCM model the information about the variations in SST makes possible the forecast of seasonal climate. The SST information may be provided to the GCM by using the SST field as a boundary condition or by coupling the GCM to an ocean model that will then provide the necessary SST information. GCM seasonal forecasts may be downscaled dynamically (e.g. Graham et al. 2007; Bastola et al. 2013; Bastola and Misra, 2014) or statistically (e.g. Uvo and Graham, 1998; Landman et al 2001; Nilsson et al. 2008), to better represent regional interests.

An early attempt to use climate model output for hydrological forecasting in a coastal Californian basin during winter 1997/1998 was made by Kim et al. (2000). They found an overall decent agreement between simulated and observed discharge. Low (high) flows were however systematically overestimated (underestimated), which was attributed primarily to climate model precipitation bias. To tackle this problem of climate model biases, Wood et al. (2002) proposed bias-correction by a percentile-based mapping of the climate model output to the climatological distributions of the input variables. Recently, several investigations have focused on the relative role of uncertainties in the initial state and in the climate forecast, respectively, for the hydrological forecast skill (e.g. Li et al., 2009; Shukla and Lettenmaier, 2011).

In a climate-based statistical approach, ~~tele~~connections between climate phenomena that affects the large-scale atmospheric circulation and the subsequent hydro-meteorological development in specific locations are identified and utilised (e.g. Jónsdótir and Uvo, 2009). Such connections are known as teleconnections as they link phenomena occurring in widely separated regions of the world. The impacts of the El Niño-Southern Oscillation on the tropical climate are the most commonly use of such teleconnections in seasonal forecast (Troccoli, 2010). Teleconnections can be also the basis for seasonal forecast in high latitudes such as the ~~in~~ impacts of the North Atlantic Oscillation in the winter climate in Scandinavia (e.g. Uvo, 2003) and the more recently identified impacts of the Scandinavian Pattern on summer climate in southern Sweden (Engström, 2011; Foster and Uvo, 2012; Foster et al. 2015). Teleconnection indices have also been used as predictors in regression-based approaches to seasonal hydrological forecasting (e.g. Robertson and Wang, 2012).

In light of the above described progress of the field, it is time to explore ways of updating operational practices by incorporating the new knowledge acquired and methods developed. ~~The current spring flood forecasting practice at the Swedish Meteorological and Hydrological Institute (SMHI) is an example of the traditional model-based approach. It is a climatological ensemble approach based on the HBV hydrological model (e.g. Bergström, 1976; Lindström et al., 1997). In the procedure, HBV is initialized by running it with observed meteorological inputs (precipitation and temperature) for a spin-up period up to the forecast issue date. Then, all available historical daily precipitation and temperature series in the period from the forecast issue date to the end of the forecasting period are used as input to HBV, generating an ensemble of spring flood forecasts. The main variable delivered to end users is the median value of total accumulated discharge in the spring flood period, but also percentiles are used. While overall sound and generally useful, this current practice has the obvious limitation that it is based on the climatology, i.e. the normal climate. Thus, if the weather from the forecast issue date up to the spring flood period evolves in a close-to-normal way the median forecast is likely to have a small error. However, if the weather deviates from the climatology, the forecast error will be large.~~

The objective of this study has been to develop, test and evaluate new approaches to spring flood forecasting in Sweden. The current spring flood forecasting practice at the Swedish Meteorological and Hydrological Institute (SMHI) is an example of the traditional model-based approach. It is a climatological ensemble approach based on the HBV hydrological model (e.g. Bergström, 1976; Lindström et al., 1997). The main scientific hypothesis examined is that the application of large-scale climate data (historical and forecasted) can improve forecast skill, as compared with today's procedure. A secondary hypothesis is that a combination of approaches provides an added value, as compared with each individual approach. Three different approaches have been tested and evaluated: (1) identifying analogue historical years that resemble the weather in the current year, (2) using meteorological seasonal forecasts as input to the HBV model and (3) applying statistical relationships between large-scale circulation variables and spring flood volume.

~~- Reduced historical ensemble by analogue years. Two methods for identifying analogue years within the historical years were evaluated. Both are based on analyses of the weather development just before the forecast issue date. (1) Teleconnection indices (TCI): the evolution of different indices representing different climate phenomena. (2) Circulation patterns (CP): frequency of different groups of weather types that describe the large-scale atmospheric state.~~

~~- Meteorological seasonal forecasts as input to the dynamical hydrological model (DM). Temperature and precipitation in ensemble forecasts are converted into HBV model input.~~

~~- Statistical downscaling of accumulated discharge (SD). Statistical relationships between large-scale circulation variables and accumulated discharge are identified and calibrated for the forecast period.~~

The new approaches were evaluated for the spring flood forecasts 2000-2010 issued in January, March and May for the river~~s~~ Vindelälven~~, Ångermanälven and Ljusnan~~ in Sweden.

2 Material

2.1 Study area, local data and models

The ~~basin~~catchment~~s~~ of the river~~s~~ Vindelälven~~, Ångermanälven and Ljusnan~~ ~~have~~ has been used for testing spring flood forecast~~ (Fig. 1a)~~. Vindelälven is unregulated~~, whereas both Ångermanälven and Ljusnan are regulated. For each river basin,~~ and two stations ~~have been~~were selected for evaluation of the forecast methods; ~~one~~ Sorsele located in the upstream part of the basin and ~~one~~ Vindeln at basin outlet (Fig. 1a). ~~The upstream area ranges between 1 700 and 31 000 km² (Table 1).~~ The catchment's elevation range is ~260-840 m.a.s.l. and

~5% of the area consists of lakes. The annual mean temperature is -0.7°C and precipitation ~780 mm. Fig. 2a shows the mean hydrograph for station Vindeln (1981-2010) in the period January-July, which is the period of interest in this study. In January-February the temperature is generally below -10°C and very little runoff is generated. Melting generally starts in late April and the subsequent spring flood extends throughout July, followed by elevated discharge levels also in August-October.

In this study we focus on forecasts of the *accumulated discharge in the spring flood period* (May-July), which is the key variable delivered to the hydropower industry.~~, and t~~ This quantity will in the following be referred to as SFV (spring-flood volume). The mean SFV in ~~the study basins ranges between~~ station Vindeln (Table 1)~~m³~~, ~~corresponding~~ corresponds to an average discharge~~s~~ in the spring flood period ~~between approximately 100 and 1000~~of ~ 380 m³/s. SFV has a pronounced inter-annual variability, which is illustrated by its range (Table 1) and frequency distribution (Fig. 2b).

The HBV model (Bergström, 1976; Lindström et al., 1997) was set up and calibrated for ~~all three rivers~~Vindelälven~~.~~, divided into 18 sub-catchments with a mean size of 740 km². HBV is a rainfall-runoff model which includes conceptual numerical descriptions of hydrological processes at basin scale. The general water balance in the HBV model can be expressed as

$$P - E - Q = \frac{d}{dt}\left[SP + SM + UZ + LZ + VL\right] \qquad (1)$$

where P denotes precipitation, E evapotranspiration, Q runoff, SP snow pack, SM soil moisture, UZ and LZ upper and lower groundwater, respectively, and ~~lakes~~ VL the volume of lake~~s volume~~. Input data are normally daily observations of P~~precipitation~~, air temperature T and monthly estimates of potential evapotranspiration; output is daily Q. ~~Air t~~Temperature ~~(T)~~ data are used for calculations of snow accumulation and melt and possibly potential evaporation. The model consists of subroutines for meteorological interpolation, snow accumulation and melt, evapotranspiration estimation, a soil moisture accounting procedure, routines for runoff generation and finally, a simple routing procedure between sub-basins and lakes. Applying the model necessitates calibration of a number of free parameters, generally about 10.

For historical simulation and calibration, daily P and T inputs for the Vindelälven basin were created from gridded fields (4×4 km²), created by optimal interpolation with altitude and wind taken into account (e.g. Johansson, 2002). These data, as well as Q observations, are available since 1961. The HBV set-up used in this experiment is the continuously updated and re-calibrated version used operationally, conceivably representing the optimal performance currently attainable. The calibration is mainly based on the historical period prior to the evaluation period (1961-1999), but some re-calibration has been done also later.

~~P and T inputs may be given either as station data or as gridded fields, and the latter are generally created by optimal interpolation (e.g. Johansson, 2002). The HBV model set-ups used here for rivers Vindelälven and Ljusnan use gridded inputs whereas the set-up for Ångermanälven uses station-based input. In all cases performed in this work, the data span was 1961 to 2010.~~

The overall accuracy of the HBV calibration ~~for each station~~ expressed in terms of the Nash-Sutcliffe efficiency (~~R^2~~NSE) and the relative volume error (RVE) in period Oct 1999 - Sep 2010 are given in Table 1. Values of ~~R^2~~NSE ~~consistently around~~ ~ 0.9 and only a few percent volume error imply an accurately calibrated model~~s~~ with limited scope for improvement.

2.2 Large-scale atmospheric data

For the definition of circulation patterns (Sect. 3.1~~2.2~~), the ERA40 data set (Uppala et al., 2005) , with resolution of 1°×1°, was used during 1961-2002 while ERAINTERIM (Dee et al., 2011), with a 0.75°×0.75° resolution, was used during 2003-2010. The domain is shown in Fig. 1a. For the teleconnection-based method studies (Sect. 3.~~2.~~1), monthly indices of the North Atlantic Oscillation, Scandinavian Pattern and East Atlantic Pattern were collected from the Climate Prediction Center (Climate Prediction Center, 2015). ~~(CPC); http://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.shtml)~~.

The atmospheric seasonal forecast data used in this work were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF). ~~The forecasts are from the System 3 that consists of an ocean analysis to estimate the initial state of the ocean, a global coupled ocean-atmosphere general circulation model to calculate the evolution of the ocean and the atmosphere, and a post-processing suite to create forecast products from the raw numerical output.~~ Two model combinations were available: the ~~Cy31r1 version of~~ ECMWF IFS (Integrated Forecast System, version 3) coupled with a 1° version of the HOPE ocean model, and the ~~Arpege~~ ARPEGE atmospheric model coupled with the variable-resolution (0.33-2°) ORCA ocean model. Atmospheric seasonal forecasts were used in two different forms;

seasonal averages from both IFS and Arpege were used in the statistical downscaling Sect. 3.~~4~~1) and daily time series from IFS were used in the dynamical modelling (Sect. 3.~~3~~1).

- Seasonal averages. These data are the ensemble means of the different predicted fields covering the domain 75°W to 75°E and 80°N to 20°N with a 2°×2° resolution. The predicted fields considered were: 2m temperature, 10m meridional wind velocity, meridional wind stress, 10m zonal wind velocity, zonal wind stress, surface sensible heat flux, surface latent heat flux, total precipitation, 850mb temperature, 850mb specific humidity, 850mb meridional wind velocity, 850mb zonal wind velocity, and 850mb geopotential height. The number of ensemble members per field is 11 for the period 1982-2006 (IFS) or 1982-2007 (Arpege) and 41 for the ~~period 2007~~ remaining years until 2010. The domain is shown in Fig. 1a.

- Daily time series. These data are the forecasted daily values of 2 m temperature and the accumulated total precipitation from the forecast issue date to the forecasting period. These data spanned a period from 2000-2010 and had a domain covering 11°E to 23°E and 55°N to 70°N with a 1°×1° resolution. ~~There were 11 ensemble members for each variable for the period 2000-2006 and 41 ensemble members for 2007-2010.~~ Fig. ~~1b~~ 1a shows this 1°×1° grid in relation to Sweden.

3 ~~Methods~~Experimental set-up

Three new approaches to seasonal hydrological forecasting are presented and compared to the current climatological ensemble procedure currently applied at SMHI: analogue ensemble, dynamical modelling and statistical downscaling. All methods are described in detail in the Supplement; below only brief outlines are given.~~this section.~~

Figure 3 shows a schematic of the “temporal set-up” of the experiments. A key issue in seasonal forecasting is the lead time (green area in Fig. 3), i.e. the period between the forecast issue date and the start of the forecasting period (blue area). It may be expected that the relative skill of the different approaches depend on the lead time. Generally, the main gain of statistical approaches is expected for long lead times. When approaching the forecasting period, the representation of the hydro-meteorological state in the HBV model becomes gradually more important and the relative skill of the current procedure is likely to increase. To assess the relative skill for different lead times, we evaluate historical forecasts (re-forecasts) issued on 1 January (1/1), 1 March (1/3) and 1 May (1/5) in the period 2000-2010.

3.1 Methods

Climatological ensemble (CE): In this procedure, HBV is initialized by driving it with observed meteorological inputs (P and T) for a spin-up period up to the forecast issue date. Then, all available historical daily P and T series in the period from the forecast issue date to the end of the forecasting period are used as input to HBV, generating an ensemble of spring-flood forecasts. See further Supplement, Sect. 1.

Analogue ensemble (AE): The hypothesis is that it is possible to identify a reduced set of historical years (an analogue ensemble) that describes the weather in the coming forecasting period better than the full historical ensemble used in CE. Two methods for identifying analogue years are used, both based on analyses of large-scale atmospheric conditions 1-6 months prior to the forecast issue date (Fig. 3). (1) Teleconnection indices (TCI): evolution of indices representing different climate phenomena. (2) Circulation patterns (CP): frequencies of weather types that describe the large-scale atmospheric state. The analogue ensemble is then used in the same way as the full ensemble in the CE method. See further Supplement, Sect. 2.

Dynamical modelling (DM): HBV is initialized as in the CE method. Then T and P from meteorological seasonal forecasts (Sect. 2.2) are converted to HBV input and used to drive the model in the forecasting period. See further Supplement, Sect. 3.

Statistical downscaling (SD): Statistical relationships between forecasted large-scale circulation variables (predictors) and SFV (predictand) are identified. The predictors are defined in the 3-month period following the forecast issue date (Fig. 3). See further Supplement, Sect. 4.

~~3.1~~3.2 ~~Experimental set-up and e~~Evaluation

~~A key issue in seasonal forecasting is the lead time, i.e. the period between the forecast issue date and the start of the forecasting period. It may be expected that the relative skill of the different approaches depend on the lead time. Generally, the main gain of statistical approaches is expected for long lead times. When approaching the forecasting period, the representation of the hydro-meteorological state in the HBV model becomes gradually more important and the relative skill of the current procedure is likely to increase. To assess the~~

~~relative skill for different lead times, we evaluate hindcasts issued on the 1st of January (1/1), 1st of March (1/3) and 1st of May (1/5) in the period 2000-2010.~~

As described in the Supplement, all methods generate ensemble forecasts (although the AE approach may become deterministic if only one analogue year is found). The ensemble size, however, varies between methods as well as between years for the same method (Supplement, Table 1). Although probabilistic forecasts are generally more useful than deterministic ones, for this initial assessment, with only an 11-year evaluation period, we consider it sufficient with a deterministic evaluation. Thus, from all ensemble forecasts the median forecast is calculated and used in the subsequent analysis, neglecting any impact of ensemble size on the skill of the median (e.g. Buizza and Palmer, 1998).

~~In the SD procedure, the average circulation fields forecasted by the GCMs for the 91 days following the forecast issue date were used as predictors. It is expected that the approximation to the spring flood period improves the GCM forecast skill. The predictor fields were different for different forecast issue dates. They were selected by an initial screening based on previous literature followed by an analysis of predictive skill in the historical period.~~

Forecast performance is assessed by $MARE_F$, the mean absolute value of the relative error of a certain forecast (or simulation) F, defined as

$$MARE_F = \frac{1}{11}\sum_{y=2000}^{2010} ARE_F^y \tag{6}$$

where y denotes year and ARE_F^y the absolute value of the relative error

$$ARE_F^y = \left| 100 * \left(\frac{SFV_F^y - SFV_{OBS}^y}{SFV_{OBS}^y} \right) \right| \tag{7}$$

where OBS denotes observation.

To quantify the gain of the new forecast approaches (Sects. 3.2-3.4), their MARE-values are compared with the MARE obtained using the current CE procedure ($MARE_{CE}$) by calculating the relative improvement RI (%) according to

$$RI_F = 100 * \left(\frac{MAE_{CE} - MAE_F}{MAE_{CE}} \right) \tag{8}$$

where a positive RI indicates that the error of the new approach is smaller than the error in the CE procedure, and vice versa, and RI=100% implies a perfect forecast.

As an additional performance measure, we ~~also calculate~~use the frequency of years FY^+ (%) in which the new approach performs better (i.e. has a lower ARE) than the CE procedure. This may be expressed as

$$FY_F^+ = 100 * \left(\frac{1}{11} \sum_{y=2000}^{2010} H^y \right) \tag{9}$$

where H is the Heaviside function defined by

$$H^y = \begin{cases} 0, AE_{CE}^y < AE_F^y \\ 1, AE_{CE}^y > AE_F^y \end{cases} \tag{10}$$

As expected considering the short 11-year evaluation period, MARE is sensitive to single years with a high ARE-value. As shown in the results below (Sect. 4), in several cases this makes RI negative even if the new approach outperforms CE in most years (i.e. FY^+>50). Thus, in this study we consider FY^+ to be the most relevant measure of forecast performance, although in practice this should be determined together with end-users of the forecasts, based on e.g. the impacts of very inaccurate forecasts.

3.3 Baseline simulations with climatological ensemble (CE)

~~4 Baseline simulations with climatological ensemble (CE)~~

Before testing the new forecasting approaches, the performance of HBV model and the climatological ensemble procedure (CE) was assessed (Table 1). In simulation mode, i.e. using the actually observed values of P and T in each year, the MARE of SFV ~~varies from 4.1% in Kultsjön to 10.3% in Dönje with an average of 7.7% for all rivers~~is 7-8%. This quantifies the HBV model error and corresponds to having a perfect meteorological forecast. ~~It may be noted that the station with the lowest HBV performance in terms of the overall measures R² and RVE, Kultsjön (Table 1), in fact shows the best performance with respect to the estimated spring flood volumes (MAE=4.1%; Table 2). This difference is not a contradiction, as MAE here represents performance in one single season, and it underlines the need to complement overall calibration criteria with season-specific measures for tailored forecasting models.~~

In CE forecast mode, i.e. using P and T from all historical years as input and calculate the median SFV, the average MARE decreases gradually from ~ 20~~21.9~~% in the 1/1-forecasts to ~9~~13.4~~% in the 1/5-forecasts~~ (Table 2)~~, which thus quantifies the improvement when approaching the spring flood period. ~~Overall, the forecast accuracy decreases from north to south. This is likely related to the higher probability of having melting episodes before the spring flood in the southern part of the region considered, so that part of the accumulated snow during winter has already melted and infiltrated when the spring flood starts. The occurrence and (non-linear) effects of such early melting episodes are very difficult to accurately simulate and forecast. It is further surprising that the skill of the 1/3-forecasts in Ljusnan is slightly lower than that of the 1/1-forecasts. Conceivably, the fact that observed P and T for Jan-Feb are used as inputs to the 1/3-forecast should improve the forecasts as compared to using a climatological input ensemble for estimating the initial conditions, as is the case for the 1/1-forecast. As the difference in skill is small, we assume that the apparent illogicality is a function of the limited sample size and the associated statistical scatter.~~

The differences in Table 1 between MARE for simulations and CE forecasts, respectively, represent the part of the total error that is related to the meteorological input. ~~On average~~In Vindelälven, this part decreases from 12.1~~14.2~~ percentage points in the 1/1-forecasts (which corresponds to ~ 60~~65~~% of the total error) to 1.8~~5.7~~ points in the 1/5-forecasts (~20~~43~~%). The relative impact of the HBV model error thus increases with decreasing lead time, which implies that the scope for improving the baseline forecasts decreases with decreasing lead time. It should be emphasised that two out of the three new forecast approaches tested here (AE and DM) aim at improving the meteorological input. They can thus only improve the forecasts in that respect; the HBV model error remains. The third method (SD), however, aims at improving total performance.

~~The relative impact of the HBV model error thus increases with decreasing lead time, which implies that the scope for improving the baseline forecasts decreases with decreasing lead time. It is remarkable that MAE for the 1/5-forecasts in Vindelälven is only slightly higher than the HBV model error. This may be interpreted as that with a proper representation of the hydro-meteorological state in the HBV model for Vindelälven on 1/5, the exact evolution of the weather in the spring flood period has only a minor impact.~~

~~Some analysis of HBV model bias was also performed, i.e. the tendency to systematically over- or underestimate SFV. In simulation mode, a small positive bias (~5%) was found with~~

~~little difference between rivers. In forecast mode, only a negligible negative bias (~1%) was found.~~

~~5~~4 Results from single methods

~~Generally, the results for different stations in the same river are similar. Therefore, the results are presented as averages over the two stations in each river.~~ An overview of the results of each approach is given in Table 2. The numbers after approaches TCI and CP correspond to the best performing version of each approach.~~, see further Sects. 5.1.1 and 5.1.2. Numbers marked in boldface indicate that the new approach performs better than the CE procedure.~~

~~5.1 Analogue ensemble (AE)~~

~~As mentioned in Sect. 3.2~~Concerning the AE approach, both the TCI and the CP approach are based on analyses of the large-scale climatic conditions 1 to 6 months before the forecast date (see Supplement). The aim was to identify the number of months of prior climatic information, N, that generates the best performance when averaged over all forecast dates. ~~and rivers, to ensure that the selected approach is robust. For a specific forecast date and river, a different period of climatic information may perform better than the selected approach but this likely mainly reflects statistical variability in light of the rather limited sample available.~~

~~5.1.1 Teleconnection indices (TCI)~~

~~As shown in Table 4, t~~Using TCI to identify analogue years proved to be difficult and the reduced ensemble generated did generally not outperform CE for the SFV forecasts. Even the best performing TCI version, using 6 months' prior climate information (N=6; TCI6), consistently had a higher MARE than CE although it outperformed CE for most of the 11 years in station Sorsele (**Table 2**). For the 1/1-forecasts, N=6 was clearly superior but for the later forecasts N=1 and N=2 produced a similar performance.

The CP method turned out more successful and the resulting SFV forecasts on 1/1 and 1/3 for the best performing version (N=3; CP3) clearly outperformed CE in both stations (Table 2). SFV was more accurately forecasted than with CE in ¾ of all years. For the 1/5-forecasts, however, CP was less accurate than CE. For the 1/1- and 1/3-forecasts, N=3 was clearly superior but for the 1/5-forecasts N=2 and N=4 performed slightly better.

~~he TCI approach performs better than CE in only a few cases. The accuracy of the TCI forecasts in Vindelälven and Ljusnan is generally low. In Ångermanälven, however, the TCI forecasts are notably better and even slightly better than CE when averaged over all dates and TCI versions (i.e. number of months used). In particular, the TCI 1/5- forecasts are clearly better than the CE ones. The main reason for this difference lies in the physics that support the TCI method. This method is based on the effect of different climate phenomena on T and P and consequently discharge, and this effect varies depending on the location of the river basin (see Uvo, 2003). In particular, Ångermanälven is located in a region that is more affected by natural climate phenomena than Vindelälven and Ljusnan. It may be remarked that the different TCI versions often identify approximately the same analogue years, therefore the performance is generally rather similar for a certain forecast issue date and river.~~

~~On average, the TCI forecasts generally have a 10-20% larger MAE than CE. The best overall performance is found for TCI6, with a 5.7% larger MAE than CE. It outperforms CE in only one case but is always close to the CE accuracy. The other TCI versions (1 to 5) outperform CE slightly in few cases, but have a substantially larger error than CE in many cases.~~

~~Circulation patterns (CP)~~

~~As shown in Table 5, comparing the different CP versions (1 to 6), using a period of three months before the forecast date (CP3) to characterise the climate stands out as the superior choice. The MAE of the CP3 forecasts are on average 1.6% lower than CE and the performance gradually decreases for both shorter and longer periods. On average in Vindelälven and Ångermanälven, CP3 performs 7.3 % better than the CE forecasts and in these rivers the CP3 approach outperforms CE on essentially all forecast dates. The only exception is the 1/5-forecast for Vindelälven, which was previously shown very difficult to improve by changing the meteorological input (see discussion in connection with Table 2). The most notable improvement is found for the 1/1- and 1/3-forecasts in Vindelälven, for which the MAE is reduced by 10-25% compared with CE and the CP3-forecast is better for 75% of the forecasts used in the testing. Also for Ångermanälven, the CP6 forecast is generally better than CE with a MAE reduction of up to 25%. If considering only the meteorological input error, the average improvement by CP3 is ~30%. The relatively poor performance of the CP approach in Ljusnan is likely at least partly because the CPs were not optimised for climate around the Ljusnan, thus the local meteorological characteristics are not well described.~~

~~As mentioned in Sect. 3.1.2, the circulation patterns were defined using the ERA40 analysis and then applied to the ERAINTERIM analysis to obtain results for 2003-2010. This implies a higher uncertainty in the results for 2003-2010. If considering only the results for 2000-2002, in which the selection of analogue years is fully consistent with the CP classification, the accuracy of the CP3 forecasts improves by 10-20% as compared with the results in . This result should be interpreted with care in light of the very limited sample used, but it indicates that improved performance is attainable if using a consistent data set for the CP classification.~~

~~5.2 Dynamical modelling (DM)~~

Overall, the DM approach of using ECMWF seasonal forecasts of T and P as inputs to the HBV model did not improve performance as compared with the CE procedure (Table 2). In total, a similar performance to CE was found in station Sorsele but the accuracy in station Vindeln was consistently lower. In the 1/5-forecasts, however, DM is the overall best performing new approach. ~~Even though the DM forecasts do outperform CE in about half of the cases, on average, their MAE is higher than the CE ones for all forecast dates and rivers.~~

~~To understand why better performance was not attained, T and P from the ECMWF seasonal forecasts were compared with observations from the river basins. The results are overall similar for Vindelälven and Ljusnan. A substantial positive bias is evident for P in late winter and early spring (February-April), up to 75%, in both the 1/1- and the 1/3-forecasts. In the 1/5-forecasts, also the May P is clearly overestimated. In July, a clear negative bias is found on all forecast dates. The T bias is generally small in the period January-May, but a distinct positive bias is found in summer (June-July). Further, the seasonal forecasts become consistently warmer the closer to the spring flood period they are issued. It may be mentioned that a new version of the ECMWF seasonal forecasting system has been released. A quick look on data from the new system, which became available by the time of writing this manuscript, indicated a similar P bias but distinctly improved T-forecasts with only a small bias in the summer.~~

~~5.3 Statistical downscaling (SD)~~

~~The finally identified predictors were (with forecast issue dates in parentheses): 2m temperature (1/1, 1/3, 1/5); 850mb temperature (1/1, 1/3, 1/5); total precipitation (1/1, 1/3, 1/5); surface sensible heat flux (1/1, 1/3, 1/5); surface latent heat flux (1/1, 1/3); meridional~~

~~wind stress (1/1, 1/5); 10m meridional and zonal wind velocity (1/3, 1/5); 850mb specific humidity (1/3, 1/5); zonal wind stress (1/3, 1/5); 850mb zonal wind velocity (1/1); 850mb zonal wind velocity and geopotential height (1/3).~~

The SD method outperformed CE in the 1/1-forecasts with an RI of almost 20% in both stations (Table 2). For the 1/3- and 1/5-forecasts the SD method has FY^{+}-values > 50 in station Sorsele but RI-values of ~ -65%. This implies that the SD-forecast is generally better than CE but that it may also be very wrong.

~~The SD method only performs better than CE in a limited number of occasions, notably for the 1/1-forecasts where the SD method performs significantly better than CE on average for Vindelälven and Ångermanälven (Table 3). The likely reason for the SD method not improving on the 1/1-forecasts for Ljusnan is its location further south than the other basins and the associated complexity of the spring flood process (see Chap. 4). Even though the SD method does sporadically perform better than CE for the other forecast issue dates, these are intermingled with forecasts where the SD method performs notably worse than CE.~~

The performance of the SD method is heavily affected by whether the climatic features in the forecasting data were encountered in the training period dataset. If the forecasted conditions are outside the range encountered~~scope~~ in the training period, the SD method has the tendency to produce forecasts that differ drastically from the observations. This can be dealt with by either increasing the length of the training dataset or by analysing the year in question and determining if there were similar years in the training period which would give an indication as to how the method might perform.

With very few exceptions, the new approaches performed better in the upper part of the catchment (Sorsele) than in the outlet (Vindeln). This has not been analysed in any depth, but it is likely related to the more clear-cut spring flood in the upper part with very little prior runoff. In the outlet, melting episodes before the spring flood onset lead to temporary increased runoff and a reduction of the snow pack. These episodes, and their impacts, are likely very difficult to capture in seasonal forecasts.

~~6~~5 Composing a multi-~~model~~ method system

A multi-~~model~~ method forecast approach consists in combining forecasts resulting from different ~~models~~ methods to reach a more reliable estimate of the forecast probability distribution. This technique has been used since early 1990s for developing seasonal climate forecast (Tracton and Kalnay, 1993) and has proved to provide more skilful results than a simple model forecast (Hagedorn et al., 2005; among many others).

There are many possible ways of combining or merging multi-~~model~~ method forecasts, ranging from simple rank-based methods to more sophisticated statistical concepts. In light of the limited material available in this study, we restricted ourselves to testing two conceptually straight-forward ways of combining the forecasts: a median approach (Sect. ~~6~~5.1) and a weighted approach (Sect. ~~6~~5.2). Further, the value of using transparent and easily communicated approaches should not be underestimated when the target is operational forecasting and its associated end-user interaction.

In each approach, two method ensembles are tested. The first ensemble, denoted NEW, represents the new approaches to spring flood forecasting considered in the study and thus includes approaches AE, DM and SD. As only one approach to analogue ensemble generation should be included, the best performing one for each forecast date was used, i.e. CP for 1/1 and 1/3 and TCI for 1/5 (Table 2). The CP method is, however, not directly applicable in operational forecasting as it is based on ERA reanalyses that are only available with a time lag of several months. Further, the TCI approach does not outperform CE in the 1/5-forecasts. Therefore we also consider a second ensemble that represents what is attainable operationally. In this ensemble, denoted OPE, AE is replaced by CE and thus no attempt to identify analogue years is made here.

~~The multi-model forecast is composed of both the baseline forecast (CE) and the ones resulting from the four new approaches, including the best performing versions of the AE models (TCI6 and CP3). If any of the new methods could not generate any forecast, it was replaced by CE (e.g., the CP approach was replaced by CE when the selection algorithm could not find any analogue years).~~

~~6.1~~5.1 Median multi-~~model~~method

~~The motivation for using the median of all forecast methods is that the final result will be less influenced by extreme high or low forecasts, when compared to calculating a mean forecast.~~ As ~~five~~ three forecast are available, the median approach amounts to using the ~~third~~ second member in the ranked forecast ensemble. For the NEW ensemble, RI is indicates a clear improvement in the 1/1-forecasts as compared with CE, but no improvement in terms of FY$^+$.

The 1/3-forecasts are better than CE 60% of the time and MARE is slightly reduced on average. The 1/5-forecasts are slightly better than CE in Sorsele but slightly worse in Vindeln. On average, a slight improvement over CE is found. In the OPE ensemble, the 1/1-forecasts perform slightly better than the NEW ensemble but the 1/3-forecasts clearly worse, as expected from the good performance of CP in these forecasts (Table 2). Overall the performance of the OPE ensemble is very similar to the NEW ensemble.

In total, a reduction of MARE by up to 25% appears attainable for the 1/1-forecasts by the median approach. At the later forecast issue dates, a limited improvement in terms of both RI and FY+ was attained for Sorsele but not for Vindeln. Over all forecast dates and stations, a slight improvement over CE is indicated.

~~The average RI of the median approach is 3.6% (Table 6). Interestingly, especially for the 1/1- but also for the 1/5-forecasts the performance in Ljusnan is consistently better than any single forecast (for the 1/3-forecast CE and DM are slightly better). This demonstrates the potential gain of the multi-model approach. Also for the 1/3-forecasts in Ångermanälven the median outperforms all single forecasts. Generally for Vindelälven and Ångermanälven, one of the single forecasts outperforms the median. It may be concluded that the potential improvement from the median multi-model approach is rather limited in size, up to about 10% compared with CE for single dates and rivers, but also rather stable.~~

~~6.2~~5.2 Weighted multi-~~model~~method

This approach consists of applying weights w between 0 and 1 to the different forecasts and then adding them together. The spring flood volume forecasted by the weighted multi-~~model~~method, SFV_{FW}, is thus defined as

$$SFV_{FW} = \sum_{f=1}^{3} w_f \cdot SFV_f \quad \text{with} \quad \sum_{f=1}^{3} w_f = 1 \text{ and } w_f \geq 0 \qquad (11)$$

where the index f refers to the three different forecast methods available in each of the ensembles NEW and OPE.~~(f=1,…,N where N=5 in Vindelälven and Ljusnan and N=4 in Ångermanälven where DM-forecasts were not available).~~

One set of weights are chosen for each ~~river and~~ forecast date. ~~The weighted volume is then calculated for the selected years and rivers, and averaged over these entries.~~ The selection of weights was made based on the evaluations performed in Table 2. ~~In Ljusnan and~~

~~Vindelälven, w~~With ~~five~~ three forecast methods available (in each ensemble), the best performing method ~~with the highest RI~~ (defined by considering both RI and FY$^+$) was assigned the highest weight 0.5 (3/6) ~~(0.33=5/15)~~, the second best performing method ~~with the second highest RI was assigned~~ the ~~second highest~~ intermediate weight 0.33 (2/6) ~~(0.27=4/15), and so on until~~ and the worst performing method the ~~method with the lowest RI and~~ lowest weight 0.17 (=1/6)~~(0.07=1/15)~~. ~~In Ångermanälven, with four forecasts, the weights ranged between 0.4 (4/10) and 0.1 (1/10). In both cases the weights add up to 1.~~

The weighted NEW set outperforms CE in the 1/1- and 1/3-forecasts for both stations; only the 1/5-forecasts for station Vindeln become notably better by CE. In the OPE set, similarly to the median forecast, the 1/3-forecast is notably worse than the NEW set but still with FY$^+$>50; the 1/5-forecasts are very similar. In total, weighting is not able to improve the result as compared with median approach in terms of RI. However, over all combinations of forecast dates and stations except the 1/5-forecast in station Vindeln, the weighted forecasts perform better than CE in most years. The 1/1-forecasts are better than CE in almost 2/3 of all years with a consistent MARE-reduction of 15-20% in both stations.

~~The average RI of the weighted multi-model is almost 10% (Table 6). In four out of the nine cases the multi-model forecast is better than any single forecast, and in four cases it is only slightly outperformed by one single forecast. For the 1/1-forecast in Ångermanälven, however, the combined forecast is worse than all single forecasts.~~

It should be emphasised that the same data were thus used both to estimate the weights and to assess the performance of the weighted model, as the 10-year period is too short for proper split-sample calibration and validation. Limited testing however indicated good performance of the fixed-weight approach also for independent validation data. Besides using fixed weights it was also tested to estimate optimal weights based on historical performance. This however turned out unfeasible in this study due to the limited historical data available and the associated tendency of overfitting to the calibration data.

~~7~~6 Concluding remarks

Why always best in Sorsele? Colder/higher/more easily forecasted?

On the use of multi-method: bad forecasts eliminated

~~It is clear that the current approach to spring flood forecasting in Sweden, based on the HBV model and a climatological input ensemble (CE), is overall performing on the same level as the new approaches tested.~~ None of the new approaches consistently outperformed the CE method, although improvement was indicated. The largest improvement was found for the 1/1- and 1/3-forecasts using an analogue ensemble based on circulation patterns and for the 1/1-forecasts ~~with the SD approach~~using statistical downscaling. In these cases ~~, with~~ the new approach may outperform the CE method up 75% of the time with an error reduction of ~~~30~~20%. In the 1/5-forecasts, none of the new methods clearly outperformed the CE method. ~~The largest improvement considering all forecast dates was found for the CP approach, with an error reduction of up to 25% and with up to 75% of the forecasts outperforming CE. In total, the TCI- and DM-forecasts outperformed CE in almost half of the cases, but generally the MAE was larger.~~ By combining the different methods in a multi-method, an overall slight improvement over CE was attained, with a performance for single forecast dates and stations rather close to the best performing individual method. The overall error reduction attainable by the multi-method, ~4%, may sound limited but it must be emphasised that every percent of forecast improvement potentially corresponds to large financial revenues in energy trading activities. For spring flood forecasts early in the season, particlary in January, the multi-method clearly outperformed the CE method.

~~The most promising results from the study were obtained by the multi-model approach. Using the median forecast, an improvement by ~4% was obtained with a small variation over stations and forecast dates. This improvement may sound limited but it must be emphasised that every percent of forecast improvement potentially corresponds to large financial revenues in energy trading activities. By using fixed weights based on historical performance, an even larger improvement of almost 10% was attained. More advanced ways of combining the forecasts are certainly conceivable, but the value of using transparent and easily communicated approaches should not be underestimated when the target is operational forecasting and its associated end-user~~

~~Finally,~~ It must be emphasised that these results were obtained in a preliminary feasibility study with limited data and overall basic versions of the used methods. Future studies need to include longer test periods and more stations as well as refined and better tailored versions of the forecast methods. One limitation concerns inhomogeneities of data and forecasts in the study period, e.g. the shift from ERA40 to ERA Interim in 2003 and the shift from 11 to 41 ensemble members in the seasonal forecasts in 2006/2007. A new ECMWF IFS version (4) is now available, but preliminary tests indicate a rather similar performance of SFV forecasts by the approaches concerned (DM and SD), as compared with using the version 3 data as done here. ~~The CP approach would benefit from using more consistent reanalysis data and domains better reflecting the basin-scale climate.~~ Using bias correction of the P and T input in the DM procedure would likely ~~substantially~~ improve performance, as demonstrated by e.g. Wood et al. (2002), although such pre-processing has limitations in an operational context when new model versions are released~~.~~ ~~Incorporating~~ hydrological model data, in particular snow information, in the SD method has shown promising results in preliminary tests, especially for improving the forecasts close to the spring-flood period. Development and testing along these lines are ongoing~~ and will be reported elsewhere~~.

Author contribution

C.B.U. and K.F. designed and implemented the TCI and SD approaches. W.Y. designed and implemented the CP approach. J.O. designed and implemented the DM approach and the multi-~~model~~ method composition. J.O. prepared the manuscript with contributions from mainly C.B.U. but also W.Y and K.F.

Acknowledgements

The study was performed in research projects funded by Elforsk AB and the Swedish Research Council Formas. Many thanks to Johan Södling, Jonas German and Barbro Johansson for technical assistance as well as fruitful discussion. Constructive and detailed reviews of the original manuscript by Peter Krahe, Massimiliano Zappa, Renaud Marty and Philippe Crochet are gratefully acknowledged.

References

Arheimer, B., Lindström, G., and Olsson, J.: A systematic review of sensitivities in the Swedish flood-forecasting system, Atmos. Res., 100, 275-284, 2011.

Bastola, S. and Misra, V.: Evaluation of dynamically downscaled reanalysis precipitation data for hydrological application, Hydrol. Proc., 28, 1989-2002, 2014.

Bastola, S., Misra, V., and Li, H. Q.: Seasonal hydrological forecasts for watersheds over the southeastern United States for the boreal summer and fall seasons, Earth Interact., 17, 1-22, 2013.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, SMHI Reports RHO, No. 7, Norrköping, Sweden, 1976.

Buizza, R. and Palmer, T. N.: Impact of ensemble size on ensemble prediction. Mon. Weather Rev., 126, 2503-2518, 1998.

Climate Prediction Center: http://www.cpc.ncep.noaa.gov/data/teledoc/telecontents.shtml, last access: 12 November 2015.

Day, G.: Extended streamflow forecasting using NWSRFS, J. Water Resour. Plann. Manage., 111, 157–170, 1985.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 137, 553–597, 2011.

Engström, J.: The effect of Northern Hemisphere teleconnections on the hydropower production in southern Sweden, M.Sc. thesis, Department of Earth and Ecosystem Sciences, Physical Geography and Ecosystems Analysis, Lund University, 2011.

Foster, K. and Uvo, C. B.: Regionalisation of Swedish hydrology. XXVII Nordic Hydrology Conference, Oulu, 2012.

Foster, K., Olsson, J., and Uvo, C.B.: The spatial and temporal effect of selected teleconnection phenomena on Swedish hydrology, Clim. Dyn., submitted, 2015.

Franz, K. J., Hartmann, H. C., Sorooshian, S., and Bales, R.: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin, J. Hydrometeorol., 4, 1105–1118, 2003.

Garen, D.: Improved techniques in regression-based streamflow volume forecasting, J. Water Resour. Plann. Manage., 118, 654–670, 1992.

Goddard, L., Mason, S. J., Zebiak, S. E., Ropelewski, C. F, Basher, R., and Cane, M. A.: Current approaches to seasonal-to-interannual climate predictions, Int. J. Climatol., 21, 1111–1152, 2001.

Graham, L. P., Andreasson, J., and Carlsson, B.: Assessing climate change impacts on hydrology from an ensemble of regional climate models, model scales and linking methods - a case study on the Lule River basin, Climatic Change, 81, 293-307, 2007.

Hagedorn, R., Doblas-Reyes, F., and Palmer, T.: The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept, Tellus A, 57, 219-233, 2005.

Johansson, B.: Estimation of areal precipitation for hydrological modelling, Ph.D. thesis, Earth Sciences Centre, Göteborg University, Report nr. A76., 2002.

Jónsdóttir, J. F. and Uvo, C.B.: Long term variability in Icelandic hydrological series and its relation to variability in atmospheric circulation over the North Atlantic Ocean, Int. J. Climatol., 29, 1369-1380, 2009.

Kim, J., Miller, N. L., Farrara, J. D., and Hong, S.-Y.: A seasonal precipitation and stream flow hindcast and prediction study in the western United States during the 1997/98 winter season using a dynamic downscaling system. J. Hydrometorol., 1, 311-329, 2000.

Landman, W. A., Mason, S. J., Tyson, P. D., and Tennant, W. J.: Statistical downscaling of GCM simulations to streamflow, J. Hydrol., 252, 221-236, 2001.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, J. Geophys. Res., 114, D04114, doi:10.1029/2008JD010969, 2009.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272-288, 1997.

Nilsson, P., Uvo, C. B., and Berndtsson, R.: Monthly runoff simulation: comparison of a conceptual model, neural networks and a combination of them, J. Hydrol., 321, 344-363, 2006.

Nilsson, P., Uvo, C. B., Landman, W., and Nguye, T. D.: Downscaling of GCM forecasts to streamflow over Scandinavia, Hydrol. Res., 39, 17-25, 2008.

Pagano, T. C., Garen, D. C., Perkins, T. R., and Pasteris, P. A.: Daily updating of operational statistical seasonal water supply forecasts for the western U.S., J. Am. Water Resour. As., 45, 767–778, 2009.

Robertson, D. E. and Wang, Q. J.: A Bayesian approach to predictor selection for seasonal streamflow forecasting, J. Hydrometeorol., 13, 155–171, 2012.

Rosenberg, E. A., Wood, A. W., and Steinemann, A. C.: Statistical applications of physically based hydrologic models to seasonal streamflow forecasts, Water Resour. Res., 47, W00H14, doi:10.1029/2010WR010101, 2011.

Shukla, S. and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529-3538, doi:10.5194/hess-15-3529-2011, 2011.

Tracton, M. S. and Kalnay, E.: Operational ensemble prediction at the National Meteorological Center: practical aspects, Weather Forecast., 8, 379-398, 1993.

Troccoli, A.: Seasonal climate forecasting, Meteorol. Appl., 17, 251–268, 2010.

Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis. Q. J. Roy. Meteor. Soc., 131, 2961-3012, 2005.

Uvo, C. B.: Analysis and regionalization of northern European winter precipitation based on its relationship with the North Atlantic Oscillation, Int. J. Climatol., 23, 1185-1194, 2003.

Uvo, C. B. and Graham, N.: Seasonal runoff forecast for northern South America: a statistical model, Water Resour. Res., 34, 3515-3524, 1998.

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D.: Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res., 107, 4429, doi:10.1029/2001JD000659, 2002.

~~Arheimer, B., Lindström, G., and Olsson, J.: A systematic review of sensitivities in the Swedish flood-forecasting system, Atmos. Res., 100, 275-284, 2011.~~

~~Bárdossy, A., Stehlík, J., and Caspary, H.: Automated objective classification of daily circulation patterns for precipitation and temperature downscaling based on optimized fuzzy rules, Clim. Res., 23, 11–22, 2002.~~

~~Barnston, G. and Livezey, R. E.: Classification, seasonality and low-frequency atmospheric circulation patterns, Mon. Weather Rev., 115, 1083-1126, 1987.~~

~~Barry, R. G. and Perry, A. H.: Synoptic climatology, methods and applications, Methuen, London, U.K., 1973.~~

~~Bastola, S. and Misra, V.: Evaluation of dynamically downscaled reanalysis precipitation data for hydrological application, Hydrol. Proc., 28, 1989-2002, 2014.~~

~~Bastola, S., Misra, V., and Li, H. Q.: Seasonal hydrological forecasts for watersheds over the southeastern United States for the boreal summer and fall seasons, Earth Interact., 17, 1-22, 2013.~~

~~Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, SMHI Reports RHO, No. 7, Norrköping, Sweden, 1976.~~

~~Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns on climate data. J. Clim., 5, 541–560.~~

~~Cheng, X., and Dunkerton, T. J.: Orthogonal rotation of spatial patterns derived from singular value decomposition analysis, J. Clim., 8, 2631-2643, 1995.~~

~~Comas-Brua, L. and McDermott, F.: Impacts of the EA and SCA patterns on the European twentieth century NAO–winter climate relationship, Q. J. Roy. Meteor. Soc., 140, 354–363, 2014.~~

~~Day, G.: Extended streamflow forecasting using NWSRFS, J. Water Resour. Plann. Manage., 111, 157–170, 1985.~~

~~Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteor. Soc., 137, 553–597, 2011.~~

~~Engström, J.: The effect of Northern Hemisphere teleconnections on the hydropower production in southern Sweden, M.Sc. thesis, Department of Earth and Ecosystem Sciences, Physical Geography and Ecosystems Analysis, Lund University, 2011.~~

~~Foster, K. L. and Uvo, C. B.: Seasonal streamflow forecast: a GCM multi-model downscaling approach, Hydrol. Res., 41,503-507, 2010.~~

~~Foster, K. and Uvo, C. B.: Regionalisation of Swedish hydrology. XXVII Nordic Hydrology Conference, Oulu, 2012.~~

~~Foster, K., Olsson, J., and Uvo, C.B.: The spatial and temporal effect of selected teleconnection phenomena on Swedish hydrology, Clim. Dyn., submitted, 2015.~~

~~Franz, K. J., Hartmann, H. C., Sorooshian, S., and Bales, R.: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin, J. Hydrometeorol., 4, 1105–1118, 2003.~~

~~Garen, D.: Improved techniques in regression-based streamflow volume forecasting, J. Water Resour. Plann. Manage., 118, 654–670, 1992.~~

~~Goddard, L., Mason, S. J., Zebiak, S. E., Ropelewski, C. F, Basher, R., and Cane, M. A.: Current approaches to seasonal-to-interannual climate predictions, Int. J. Climatol., 21, 1111–1152, 2001.~~

~~Graham, L. P., Andreasson, J., and Carlsson, B.: Assessing climate change impacts on hydrology from an ensemble of regional climate models, model scales and linking methods – a case study on the Lule River basin, Climatic Change, 81, 293-307, 2007.~~

~~Hagedorn, R., Doblas-Reyes, F., and Palmer, T.: The rationale behind the success of multi-model ensembles in seasonal forecasting I. Basic concept, Tellus A, 57, 219-233, 2005.~~

~~Hamlet, A. and Lettenmaier, D.: Columbia River streamflow forecasting based on ENSO and PDO climate signals, J. Water Resour. Plann. Manage., 125, 333–341, 1999.~~

~~Hay, L., McCabe, G., Wolock, D., and Ayers, M.: Simulation of precipitation by weather type analysis, Water Resour. Res., 27, 493–501, 1991.~~

~~Hurrell, J. W. and Deser, C.: North Atlantic climate variability: The role of the North Atlantic Oscillation, J. Marine Syst., 79, 231–244, 2010.~~

~~Johansson, B.: Estimation of areal precipitation for hydrological modelling, Ph.D. thesis, Earth Sciences Centre, Göteborg University, Report nr. A76., 2002.~~

~~Johnson, S.: Hierarchical clustering schemes, Psychometrika, 32, 141–154, 1967.~~

~~Jónsdóttir, J. F. and Uvo, C.B.: Long term variability in Icelandic hydrological series and its relation to variability in atmospheric circulation over the North Atlantic Ocean, Int. J. Climatol., 29, 1369-1380, 2009.~~

~~Kim, J., Miller, N. L., Farrara, J. D., and Hong, S.-Y.: A seasonal precipitation and stream flow hindcast and prediction study in the western United States during the 1997/98 winter season using a dynamic downscaling system. J. Hydrometorol., 1, 311-329, 2000.~~

~~Kushnir, Y.: Europe’s winter prospects, Nature, 398, 289-291, 1999.~~

~~Kyselý, J., and Huth, R.: Changes in atmospheric circulation over Europe detected by objective and subjective methods. Theor. Appl. Climatol., 85, 19-36, 2005.~~

~~Landman, W. A., Mason, S. J., Tyson, P. D., and Tennant, W. J.: Statistical downscaling of GCM simulations to streamflow, J. Hydrol., 252, 221-236, 2001.~~

~~Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, J. Geophys. Res., 114, D04114, doi:10.1029/2008JD010969, 2009.~~

~~Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272-288, 1997.~~

~~MacQueen, J.: Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, p. 281–297, 1967.~~

~~Nilsson, P., Uvo, C. B., and Berndtsson, R.: Monthly runoff simulation: comparison of a conceptual model, neural networks and a combination of them, J. Hydrol., 321, 344-363, 2006.~~

~~Nilsson, P., Uvo, C. B., Landman, W., and Nguye, T. D.: Downscaling of GCM forecasts to streamflow over Scandinavia, Hydrol. Res., 39, 17-25, 2008.~~

~~Pagano, T. C., Garen, D. C., Perkins, T. R., and Pasteris, P. A.: Daily updating of operational statistical seasonal water supply forecasts for the western U.S., J. Am. Water Resour. As., 45, 767–778, 2009.~~

~~Robertson, D. E. and Wang, Q. J.: A Bayesian approach to predictor selection for seasonal streamflow forecasting, J. Hydrometeorol., 13, 155–171, 2012.~~

~~Rosenberg, E. A., Wood, A. W., and Steinemann, A. C.: Statistical applications of physically based hydrologic models to seasonal streamflow forecasts, Water Resour. Res., 47, W00H14, doi:10.1029/2010WR010101, 2011.~~

~~Shukla, S. and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529-3538, doi:10.5194/hess-15-3529-2011, 2011.~~

~~Tracton, M. S. and Kalnay, E.: Operational ensemble prediction at the National Meteorological Center: practical aspects, Weather Forecast., 8, 379-398, 1993.~~

~~Troccoli, A.: Seasonal climate forecasting, Meteorol. Appl., 17, 251–268, 2010.~~

~~Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J.: The ERA-40 re-analysis. Q. J. Roy. Meteor. Soc., 131, 2961-3012, 2005.~~

~~Uvo, C. B.: Analysis and regionalization of northern European winter precipitation based on its relationship with the North Atlantic Oscillation, Int. J. Climatol., 23, 1185-1194, 2003.~~

~~Uvo, C. B. and Graham, N.: Seasonal runoff forecast for northern South America: a statistical model, Water Resour. Res., 34, 3515-3524, 1998.~~

~~Uvo, C.R.B., Repelli, C.A., Zebiak, S.E., Kushnir, Y.: The Influence of Tropical Pacific and Atlantic SST on Northeast Brazil Monthly Precipitation. J. Clim., 11, 551-562, 1998.~~

~~Wetterhall, F., Bárdossy, A., Chen, D., Halldin, S., and Xu, C.-Y.: Daily precipitation-downscaling techniques in three Chinese regions, Water Resour. Res., 42, W11423, doi:10.1029/2005WR004573, 2006.~~

~~Wilby, R. and Wigley, T.: Stochastic weather type simulation for regional climate change impact assessment, Water Resour. Res., 30, 3395-3403, 1994.~~

~~Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D.: Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res., 107, 4429, doi:10.1029/2001JD000659, 2002.~~

~~Yang, W., Bárdossy, A., and Caspary, H. J.: Downscaling daily precipitation time series using a combined circulation- and regression-based approach, Theor. Appl. Climatol., 102, 439-454, 2010.~~

~~Yarnal, B.: A procedure for the classification of synoptic weather maps from gridded atmospheric pressure data, Comput. Geosci., 10, 397-410, 1984.~~

~~Zadeh, L.: Fuzzy sets, Inform. Control, 8, 338-353, 1965.~~

Table 1. Basin and station characteristics including overall performance of the HBV model. MARE (%) of SFV estimated by simulation (SIM) and by climatological ensemble (CE) forecasts (F) with different issue dates (1/1, 1/3, 1/5). All values represent 2000-2010.~~Basin and station characteristics including performance of the HBV model. MARE$_{CE}$ (%) of the climatological ensemble CE simulations SIM and forecasts F with different issue dates (1/1, 1/3, 1/5) in the period 2000-2010.~~

Station	Area (km²)	HBV		SFV (m^3*10^9) Min / Mean / Max	$MARE_{SIM}$	$MARE_{CE}$		
		NSE	RVE			F 1/1	F 1/3	F 1/5
Sorsele	6 054	0.89	3.2	1.61 / 2.30 / 2.77	6.8	19.2	11.6	9.5
Vindeln*	11 846	0.91	1.5	2.26 / 3.18 / 4.11	8.2	20.0	13.2	9.0

*Basin outlet

Table 2. Relative improvement RI (%) and frequency of years with a better performance FY^+ (%) of the new forecasting approaches TCI6, CP3, DM and SD, as compared with the climatological ensemble CE (boldface indicates better performance than CE).~~Relative improvement RI (%) and frequency of years with a better performance FY+ (%) of the new forecasting approaches TCI6, CP3, DM and SD, as compared with the climatological ensemble CE. Boldface indicates that the new approach performs better than CE.~~

		TCI6		CP3		DM		SD	
		RI	FY^+	RI	FY^+	RI	FY^+	RI	FY^+
1/1	Sorsele	-6.6	**55**	**1.4**	**75**	**7.6**	45	**18.4**	**55**
	Vindeln	-9.0	45	**13.0**	**75**	**-13.5**	45	**17.3**	**55**
1/3	Sorsele	-1.2	**64**	**19.2**	**70**	-17.3	45	-63.3	**55**
	Vindeln	-10.4	45	**36.2**	**80**	-18.5	45	-29.4	45
1/5	Sorsele	-6.6	**55**	-9.9	33	**1.3**	**55**	-66.8	**64**
	Vindeln	-21.9	45	-31.3	33	-12.0	36	-90.3	27
Average		-9.3	**52**	**4.8**	**61**	-8.7	45	-35.7	50

Table 3. Relative improvement RI (%) and frequency of years with a better performance FY$^+$ (%) for the median and weighted multi-method approaches, as compared with the climatological ensemble CE (boldface indicates better performance than CE).~~Relative improvement RI (%) and frequency of years with a better performance FY$^+$ (%) for the median and weighted multi-method approaches, as compared with the climatological ensemble CE (boldface indicates better performance than CE).~~

		Median				Weighted			
		NEW		OPE		NEW		OPE	
		RI	FY$^+$	RI	FY$^+$	RI	FY$^+$	RI	FY$^+$
1/1	Sorsele	**20.9**	50	**25.3**	**56**	**20.1**	**55**	**18.2**	**64**
	Vindeln	**5.8**	50	**12.5**	**56**	**15.7**	**64**	**12.9**	**64**
1/3	Sorsele	**5.9**	**60**	-4.2	**56**	**13.3**	**64**	-7.2	**55**
	Vindeln	-0.1	**60**	-10.7	43	**3.8**	**55**	-10	**55**
1/5	Sorsele	**3.7**	**55**	**7.9**	**67**	-5.0	**55**	-0.6	**55**
	Vindeln	-15.6	36	-5.2	33	-23.3	36	-13.5	45
Average		**3.4**	**52**	**4.3**	**52**	**4.1**	**55**	0.0	**56**

Figure captions

Figure 1. Domain used in the CP method, ECMWF IFS grid (blue dots), Vindelälven catchment (yellow), stations Sorsele (S) and Vindeln (V) (a). Domain used in the SD method (b).

Figure 2. Mean annual Q cycle (a) and SFV frequency distribution (b) for station Vindeln in the period 1961-1999.

Figure 3. Temporal set-up of the experiments. Vertical black lines; forecast dates. Blue area: spring flood period. Green area: lead time. Red area: full historical period used in the selection of analogue years (CP, TCI). Black arrows: time periods (1-6 months back in time) tested in the selection of analogue years (CP, TCI). Yellow arrows: time period (3 months ahead) used to calculate the predictors in the SD method. White arrows: forecasting periods in which the HBV model was run using full historical ensemble (CE), reduced analogue ensemble (CP, TCI) and ECMWF forecasts (DM).

~~Figure 1. Locations of the three study basins (a) and the 1°x1° ECMWF grid (b).~~